Online Learning Algorithms

MATH 436 - Intermediate PDE

Robert Joseph George

Department of Mathematics and Statistics
University of Alberta

01/12/2022

Outline

Introduction

Consider learning from examples $(x_t, y_t) \in X \times \mathbb{R} (t \in \mathbb{N})$, drawn at random from a probability measure ρ on $X \times \mathbb{R}$. For $\lambda > 0$, one wants to approximate the function f_λ^* minimizing over $f \in H$ the quadratic functional

$$\int_{X \times Y} (f(x) - y)^2 d\rho + \lambda \|f\|_H^2,$$

where H is some Hilbert space. In this presentation we explore a scheme for doing this is given by using one example at a time t to update to f_t the current hypothesis f_{t-1} which depends only on the previous examples.

Main Goal

The main goal in this presentation is to showcase an online algorithm in Hilbert spaces and cover Kernel Methods. By choosing a quadratic functional to optimize one is able to give a deeper understanding of this online learning phenomenon.

Outline

Kernels

What even is a kernel?

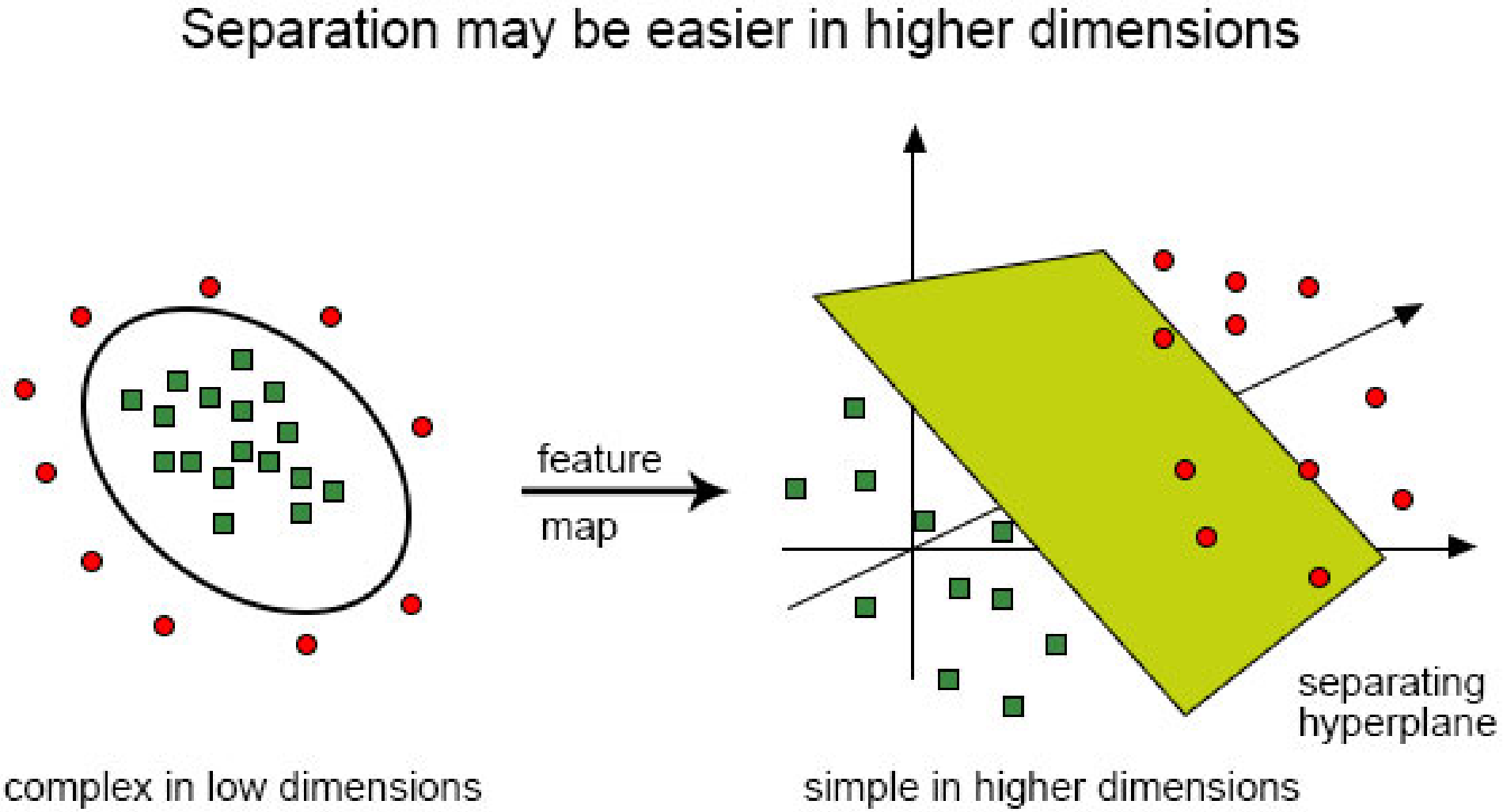
Let $\mathcal{X}$ be a nonempty set, sometimes referred to as the index set. A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive-definite (p.d.) kernel on $\mathcal{X}$ if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K\left(x_i, x_j\right) \geq 0$$

Examples

The first is the Gaussian kernel $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $K\left(x, x'\right) = \exp\left(-\left\|x - x'\right\|^2 / c^2\right)$ $(c > 0)$. The second is the linear kernel $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $K\left(x, x'\right) = \left\langle x, x'\right\rangle + 1$. The restriction of these functions on $X \times X$ will induce the corresponding kernels on subsets of $\mathbb{R}^n$.

Kernel Trick



Kernel's and PDE

The most well-known heat kernel is the heat kernel of d-dimensional Euclidean space $\mathbf{R}^d$, which has the form of a time varying Gaussian function,

$$K(t,x,y) = \exp(t\Delta)(x,y) = \frac{1}{(4\pi t)^{d/2}} e^{-\|x-y\|^2/4t} \qquad \left(x, y \in \mathbb{R}^d, t > 0\right)$$

This solves the heat equation

$$\frac{\partial K}{\partial t}(t,x,y) = \Delta_x K(t,x,y)$$

Outline

Reproducing Kernel Hilbert Space (RKHS)

Let H_K be RKHS associated with a Mercer kernel K.

How do we even construct this space?

Proof. For all x in X, define $K_x = K(x, \cdot)$. Let H_0 be the linear span of $\{K_x : x \in X\}$. Define an inner product on H_0 by

$$\left\langle \sum_{j=1}^{n} b_j K_{y_j}, \sum_{i=1}^{m} a_i K_{x_i} \right\rangle_{H_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K\left(y_j, x_i\right),$$

which implies $K(x, y) = \langle K_x, K_y \rangle_{H_0}$. The inner product is symmetric. Let H be the completion of H_0 with respect to this inner product. Then H consists of functions of the form

$$f(x) = \sum_{i=1}^{\infty} a_i K_{x_i}(x) \text{ where } \lim_{n \to \infty} \sup_{p \geq 0} \left\| \sum_{i=n}^{n+p} a_i K_{x_i} \right\|_{H_0} = 0.$$

RKHS Construction (Con't)

Reproducing Property and Uniqueness

Now we can check the reproducing property:

$$\langle f, K_x \rangle_H = \sum_{i=1}^{\infty} a_i \langle K_{x_i}, K_x \rangle_{H_0} = \sum_{i=1}^{\infty} a_i K(x_i, x) = f(x).$$

To prove uniqueness, let G be another Hilbert space of functions for which K is a reproducing kernel. For every x and y in X, (2) implies that

$$\langle K_x, K_y \rangle_H = K(x, y) = \langle K_x, K_y \rangle_Q .$$

By linearity, $\langle \cdot, \cdot \rangle_H = \langle \cdot, \cdot \rangle_G$ on the span of $\{K_{ax} : x \in X\}$. Then $H \subset G$ because G is complete and contains H_0 and hence contains its completion. Then it is easy to show that every element of G is in H which implies that $f = f_H$ and concludes the proof.

Outline

Online Learning Algorithm

Algorithm in RKHS

Given a sequence of examples $z_t = (x_t, y_t) \in X \times Y (t \in \mathbb{N})$

$$f_{t+1} = f_t - \gamma_t \left(\left(f_t \left(x_t \right) - y_t \right) K_{x_t} + \lambda f_t \right), \qquad \text{for some } f_1 \in H_K, \text{ e.g. } f_1 = 0,$$

where

- for each $t \in \mathbb{N}, (x_t, y_t)$ is drawn identically and independently according to ρ,
- the regularization parameter $\lambda \geq 0$ and the step size $\gamma_t > 0$.

Outline

Quadratic Map

Consider a Hilbert space (W) with inner product $\langle , \rangle$. Consider the quadratic map $V : W \to \mathbb{R}$ given by

$$V(w) = \frac{1}{2}\langle Aw, w\rangle + \langle B, w\rangle + C$$

where $A : W \to W$ is a positive definite bounded linear operator whose inverse is bounded, i.e. $\left\|A^{-1}\right\| < \infty, B \in W$ and $C \in \mathbb{R}$. Then the gradient $\operatorname{grad} V : W \to W$ is given by

$$\operatorname{grad} V(w) = Aw + B.$$

V has a unique minimal point $w^* \in W$ such that $\operatorname{grad} V\left(w^*\right) = Aw^* + B = 0$, i.e.

$$w^* = -A^{-1}B.$$

Coercivity condition

Coercive Operators

A self-adjoint operator $A : H \to H$, where H is a real Hilbert space, is called coercive if there exists a constant $c > 0$ such that

$$\langle Ax, x \rangle \geq c\|x\|^2$$

for all x in H

Norm-Coercive mappings

A mapping $f : X \to X'$ between two normed vector spaces $(X, \|\cdot\|)$ and $(X', \|\cdot\|')$ is called norm-coercive iff $\|f(x)\|' \to +\infty$ as $\|x\| \to +\infty$

Then we get that

$$\|V(w)\| = \|\frac{1}{2}\langle Aw, w \rangle + \langle B, w \rangle + C\| \geq \|\frac{1}{2}c\|w\|^2 - \|B\|\|w\|^2 + C\|$$

Algorithm in RKHS

Our concern is to find an approximation of this point, when A, B and C are random variables on a space Z.

$$w_{t+1} = w_t - \gamma_t \operatorname{grad} V(w_t), \qquad \text{for } t = 1, 2, 3, \ldots$$

with γ_t a positive step size. For each example z, the stochastic gradient of $V_z, \operatorname{grad} V_z : W \to W$ is given by the affine map
$\operatorname{grad} V_z(w) = A(z)w + B(z)$, with $A(z), B(z)$ denoting the values of random variables A, B at $z \in Z$.

Rewritten equation

Thus the above equation then becomes: For $t = 1, 2, 3, \ldots$, let z_t be a sample sequence and define an update by

$$w_{t+1} = w_t - \gamma_t \left(A_t w_t + B_t \right), \qquad \text{for some } w_1 \in W$$

where

- $z_t \in Z (t \in \mathbb{N})$ are drawn independently and identically according to ρ;
- the step size $\gamma_t > 0$
- For each $t \in \mathbb{N}$, let $A_t = A\left(z_t\right)$ and $B_t = B\left(z_t\right)$.

Outline

Fréchet Derivative

Let V and W be normed vector spaces, and $U \subseteq V$ be an open subset of V. A function $f : U \to W$ is called Fréchet differentiable at $x \in U$ if there exists a bounded linear operator $A : V \to W$ such that

$$\lim_{\|h\| \to 0} \frac{\|f(x+h) - f(x) - Ah\|_W}{\|h\|_V} = 0.$$

The limit here is meant in the usual sense of a limit of a function defined on a metric space, using V and W as the two metric spaces, and the above expression as the function of argument h in V. As a consequence, it must exist for all sequences $\langle h_n \rangle_{n=1}^{\infty}$ of non-zero elements of V that converge to the zero vector $h_n \to 0$. Equivalently, the first-order expansion holds, in Landau notation

$$f(x+h) = f(x) + Ah + o(h).$$

If there exists such an operator A, it is unique, so we write $Df(x) = A$ and call it the Fréchet derivative of f at x.

Specific Potential Map

Consider the Hilbert space $W = H_K$. For fixed $z = (x, y) \in Z$, take the following quadratic potential map $V : H_K \to \mathbb{R}$ defined by

$$V_z(f) = \frac{1}{2}\left\{(f(x) - y)^2 + \lambda \|f\|_K^2\right\}.$$

Recall that the gradient of V_z is a map $\operatorname{grad} V_z : H_K \to H_K$ such that for all $g \in H_K$,

$$\langle \operatorname{grad} V_z(f), g \rangle_K = DV_z(f)(g)$$

where the Fréchet derivative at $f, DV_z(f) : H_K \to \mathbb{R}$ is the linear functional such that for $h \in H_K$,

$$\lim_{\|h\| \to 0} \frac{|V_z(f + h) - V_z(f) - DV_z(f)(h)|}{\|g\|} = 0.$$

Hence

$$DV_z(f)(h) = (f(x) - y)h(x) + \lambda \langle f, g \rangle_K = \langle (f(x) - y)K_x + \lambda f, h \rangle_K,$$

where the last step is due to the reproducing property $h(x) = \langle g, K_x \rangle_K$.

Proposition

Final Algorithm

Let $\operatorname{grad} V_z(f) - (f(x) - y)K_x + \lambda f$. Taking $f = f_t$ and $(x, y) = (x_t, y_t)$, by $f_{t+1} = f_t - \gamma_t \operatorname{grad} V_{z_t}(f_t)$, we have

$$f_{t+1} = f_t - \gamma_t \left(\left(f_t \left(x_t \right) - y_t \right) K_{x_t} + \lambda f_t \right),$$

which establishes the equation as shown previously. [SY06]

Outline

Thanking Remarks

Thank you to Professor Mohammad Ali Niksirat for this wonderful opportunity to present this research topic and for teaching this course. Thank you to everyone present here and for coming to my presentation. Merry Christmas and have a great winter break ahead!

References


Steve Smale and Yuan Yao, *Online learning algorithms*, Foundations of Computational Mathematics **6** (2006), 145–170.